Number Identification: A Unique Developmental Pathway in Mathematics?

Journal:	*Research Papers in Education*
Manuscript ID	RRED-2017-0213.R1
Manuscript Type:	Original Paper
Keywords:	mathematics, number identification, Rasch Measurement, learning pathways

Word Count (excluding references): 7234

Number Identification: A Unique Developmental Pathway in Mathematics?

Abstract

We make a *prima facie* case for identifying a single pathway in the learning of Hindu-Arabic numerical symbols and discuss why this ability may be a critical gateway concept in developing mathematical competencies. A representative sample of English and Scottish children were assessed using a number symbol identification paradigm in the PIPS Baseline assessment at the beginning and end of their first school year. Through a Rasch analysis of real and simulated data, we show that 1) there appears to be a single, unidimensional pathway in learning to identify number symbols with discrete difficulty stages, 2) on examination of differential item functioning, this pathway is invariant across gender, country, socioeconomic background, first language and across the first year of schooling, and 3) almost all children make progress along the pathway during the year. A number identification scale may thus be a universal ruler by which all pupils could be assessed.

Keywords: mathematics, number identification, Rasch measurement, learning pathways

Introduction

The term 'Number Sense' refers to multiple mathematical competencies including (but not exclusively) counting (1, 2, 3, 4…), magnitude (5 is more than 4 but less than 6), cardinality (final count represents total) and linear representation (3 is one more than 2 and 4 is one more than 3) (Berch, 2005; Gersten, Jordan, & Flojo, 2005; Malofeeva, Day, Saco, Young, & Ciancio, 2004; National Mathematics Advisory Panel, 2008; Siegler & Booth, 2004). On grasping these competencies, children develop more complex skills and make connections between core concepts (Gersten et al., 2005). Evidence suggests that children with difficulties in mathematics perform poorly on assessments of basic competencies (Gersten, et al., 2005; Mazzocco & Thompson, 2005).

The usefulness of the term 'Number Sense' is debatable due to its conceptual breadth. Focusing on specific mathematics facets which are demonstrable predictors of later mathematics attainment may be more useful, especially if their acquisition can be shown to develop along a single path. Studies suggest that children's ability to identify numbers (the ability to apply a number word 'two' to the Hindu-Arabic numeral '2') on entry to school consistently predicts later attainment (Chard, Clarke, Baker, Otterstedt, Braun & Katz, 2005; Clarke & Shinn, 2004; Jordan, Kaplan, Locuniak, & Ramineni, 2007; National Mathematics Advisory Panel, 2008; Tymms, 1999, Tymms, Merrell, Henderson, Albone & Jones, 2012), but much remains to be learned about how children progress in developing this competency. Gaining greater understanding of how pupils learn to identify numbers (in the context of this study, to equate a written symbol with a particular phonological representation) and in what order are important questions which may help identify the beginning of a progressive competency scale in formal mathematics, and further inform teaching and learning. We

provide *prima facie* evidence suggesting that a single pathway in number identification exists based on current literature regarding how children begin to identify and use number symbols.

Literature review

Development of numeracy skills in children

Numeracy skills development is cumulative, (Purpura & Ganley, 2014) and mathematics difficulties also tend to be cumulative when foundations are not secure, and children fall further behind (Jordan et al., 2009). Numeracy skills development typically occurs in three overlapping phases (Purpura & Ganley, 2014). Firstly, children separately learn to compare small object quantities and to count in number word sequences. Counting then develops through socially organised and structured experiences (Bertelli, Joanni & Martlew, 1998). Secondly, children apply number word sequences to fixed set object sets as well as making linkages between number words (e.g. one, two, three) and their quantities (e.g. *, **, ***). Thirdly, children develop the ability to solve story problems where, without the aid of physical objects, they combine number words and quantities to create new number words and quantities. These stages are typically called 'informal' mathematics and are often acquired prior to formal education. Additionally, to 'informal' mathematical skills, another important prerequisite for developing advanced mathematical concepts is the linking of specific number word names with their associated Hindu-Arabic numerals (e.g. 1, 2, 3 etc.). This skill does not conform to the definition of informal or formal mathematics (Baroody & Wilkins, 1999). Recently, Purpura, Baroody and Lonigan (2013) suggested that number symbol identification and the ability to understand the relationship between symbol and quantity, completely mediates the relationship between informal and later formal mathematics (each independent skill only partially mediating the relationship). This 'number knowledge' (combining identification and mapping) provides the bridge between informal number and arithmetic

knowledge to more advanced, formal mathematical protocols. Number identification ability is therefore a necessary prerequisite to further mathematical ability.

When does the ability to identify number symbols emerge? Basic numerosity can be represented non-symbolically in infants as young as 6 months (Lipton & Spelke, 2003, 2004) with symbolic representation developing from around age 3 (Gelman & Gallistel, 1978). Some evidence suggests that symbolic mapping can emerge developmentally earlier; as young as 18 months (Mix, 2009). Wynn (1992) claimed the meanings of the words "one" and "two" are learned six months apart, with the word "three" following three months later. On reaching "four" they appear to grasp the logic that each number along the scale is one more than the previous (magnitude) and that each word is uniquely associated with a specific cardinal value. The ability to map number words to symbolic and non-symbolic representations of numerosity likely develops concurrently with processes underlying the counting ability (Krajewski & Schneider, 2009; Sarama & Clements, 2009), although the use of count words such as 'one, two, three' can be unaccompanied by actual counting in very young children (Wagner & Walters, 1982). Carey (2004) suggests that direct mapping may not occur until after the cardinality principle has been grasped. Children begin to understand that numerals are distinct from other symbolic representations (i.e. letters) and are then able to map names to symbols. Around 25% of four-year olds can accurately identify numerals 1 to 9 (Ginsburg & Baroody, 2003), with some able to identify numerals 1 and 2 from as young as 18 months (Mix, 2009; Sarama & Clements, 2009). Merkley and Ansari (2016) review neurological evidence suggesting that the left intraparietal sulcus (IPS) increasingly specialises for the processing of number symbols with experience whilst the right IPS remains constant in its handling of non-symbolic representation from6 months of age. Young children know that larger single digit numerals (8, 9 etc.) are representative of larger

quantities than smaller single digits (1, 2 etc.), often independently of precise knowledge about meaning (Le Corre & Carey, 2007). As children make associated mappings to number symbols, they may use early ones as anchors to support the rapid learning of further symbols and number words (Lipton & Spelke, 2005; Mix, Prather, Smith & Stockton, 2014).

Children in the early years are capable of identifying and to some extent, understanding, multi-digit numbers (Mix et al., 2014). Mix et al., demonstrated that children as young as three and a half years old could identify single, double, triple and quadruple digit numbers successfully when presented with number pairs (e.g. 2 vs 8). Single and double-digit items were identified with greater accuracy. Some children are thus able to successfully identify Hindu-Arabic numerals of increasing complexity without formal instruction. Mapping spoken multi-digit numbers to numerals is difficult (Byrge, Smith & Mix, 2014; Fuson, 1990) and possibly linked to understanding place value. It involves reconciling morphemic representations of unit size (multiunit names) with written relative position indicators and, knowing that zero, while symbolically distinct, is not used verbally. Children must also learn associations between the first nine numerals, and that numerals positioned increasingly to the left denote increasing base-10 values. More than nine of a given value can be verbalised, but not written in symbols, as values greater than nine move to the next value on the left. Dissociation between symbols and words is also confounded by both linguistic irregularities (-teen and -ty) and that the size of numbers increases from right to left rather than, as with reading, left to right (Fuson, 1990). The development of place value (and thus the identification of multi-digit numbers) is therefore complex. Mix et al., demonstrate that preschool children have a partial understanding of place value, and can use such knowledge to interpret multi-digit numbers of increasing complexity. Performance on multi-digit tasks increased with age and grade, and ceiling effects did not emerge until second grade, where

formal place value instruction will have been. Children may thus use informal place value knowledge to expand their repertoire of digits by making inferences about numerals. This may explain how children progress from single to multi-digit numbers of increasing magnitude.

Mix et al., suggested that early learning of number symbols and place value notation happens informally through exposure within developmental environments, which naturally contains complex numeral related stimuli (building blocks, parental play, room numbers, phone numbers etc.). Natural exposure to numbers may also reinforce why numbers "one, two, three…" are learned in this order. Benford's Law (1938) suggests that in natural data sets, numbers decrease in frequency with increases in magnitude. Numbers beginning with 1 appear approximately 30% of the time, 2 approximately 17% of the time, up to 9 which appears approximately 4% of the time. This highlights the more common usage and exposure of lower digit numbers within social and developmental environments. Research indeed suggests that certain written 'benchmark' numbers appear more often in the developmental environment eg.10, 100 etc. (Byrge et al., 2014; Dehaene & Mehler, 1992). These numbers are possibly learned earlier because they are common (to children and the parents/educators that interact with them) within formal and informal instruction and play.

Number and language

Mix, Huttenlocher and Levine (2002) suggest that unlike other words, number words are novel due to the additional symbolic representations of numerals (i.e. 1, 2, 3…) between spoken and written forms. However, number words up to and including nine each have one distinct spoken, written and symbolic values and that the transparency of these early values likely makes associations between them simple. This contrasts to learning words, where the

print to speech relationship is more complex. LeFevre et al., (2010) showed that a number identification task correlated significantly and positively to measures of phonology and vocabulary in a longitudinal study of 182 children from kindergarten to grade 2. They suggest that early linguistic skills are a precursor to the early symbolic number system, subsequently leading to further mathematical understanding. Children with numeracy difficulties often also have language and literacy difficulties (Purpura & Reid (2016).

The difference between cultures in the acquisition of number words is often attributed to language. Seron and Fayol (1994) compared Belgian (Walloon) and French speaking children. In Walloon, multi-digit decade numbers are regular (70 is septante) but irregular in French (70 is soixante-dix). Seron et al., noted that Walloon speakers were more accurate on transcoding tasks involving decades than French speaking equivalents. Pixner, Zuber, Hermanova, Kaufmann, Nuerk and Moeller (2011) review evidence of similar inversion errors occurring across multiple language groups. Although limited by cultural confounds, Pixner et al., examined the effects of language structure on transcoding within the Czech language, unique for its two different number-word systems (one with inverted decades, the other without). Seven-year-old Czech speakers made more errors on the inverted than non-inverted number system. Miller, Kelly and Zhou (2005) found that Chinese speaking pre-school children develop their ability to count to 100 earlier than English speaking pre-school children and ascribe this to the base-ten structure within Chinese number words (e.g. 10-1, 10-2, 10-3, represent eleven, twelve, thirteen respectively (Zhou, 2006; Miller, Smith, Zhu & Zhang, 1995; Miller et al., 2005)), compared to the irregular English number naming system. However, Miller et al. (1995), asked pre-school children to count as high as possible but were not required to recognise written numbers. Each decade boundary showed significant increases in difficultly for both Chinese and English-speaking children. The aggregation of

this evidence suggests the development of numerical cognitions (including number symbol identification) may be contingent on linguistic structures of number systems.

Predictors of later attainment

The ability to identify numbers and letters on entry to school (aged 4) is a good predictor of later attainment (Chard et al., 2005, Clarke & Shinn, 2004; Jordan et al., 2007, National Advisory Panel, 2008; Tymms, 1999, Tymms, et al., 2012). The US National Mathematics Advisory Panel (2008) reported that children's mathematical knowledge on entry to Kindergarten and First Grade predicted achievement throughout their school career. Tymms (1999) found a correlation of 0.61 between number identification on entry to formal schooling in the UK and mathematics in year 2 and 0.60 for letter identification at the start of school and mathematics in year 2. Martin, Cirino, Sharp and Barnes (2014) found that while number identification and symbol comparison were strongly correlated with counting skills during Kindergarten, symbolic number skills predicted a greater proportion of variance in a later First Grade maths test.

The importance of early number competence (understanding the meaning of number words, symbols and number relationships) on learning trajectories is detailed by Purpura et al., (2014) and Jordan et al., (2009). Children demonstrating higher number competence in Kindergarten showed a significantly greater performance in Third Grade mathematics and a modest but significant increased rate of achievement. Jordan et al., (2009) also show how number sense in Kindergarten (defined as; counting, number knowledge, nonverbal problems, story problems and number combinations) is a significant predictor of children's ability to solve applied mathematics problems in both First and Third Grade. Research into

early mathematical disabilities also indicates that affected children often have specific difficulties with number symbols rather than with informal mathematical processes or other cognitions (Butterworth & Regiosa, 2007; Rousselle & Noel, 2007; Song & Ginsburg, 1987).

Teaching effects

Chard et al., (2005) demonstrated that during kindergarten, number identification ability showed greater progress compared with other mathematical competencies over the first 36 teaching weeks. Mix et al., (2014) also showed that number identification accuracy increased between kindergarten and second grade and that performance could be boosted with explicit symbol training. These studies suggest that children progress rapidly in identifying numbers once formalised learning begins.

Children's number symbol knowledge on entry to education is dependent upon the input they have received (Jordan et al., 2009). Evidence suggests that number identification abilities show high levels of variability in preschool children (Mussolin, Nys, Content & Leybaert, 2014). Research into how teaching ameliorates gaps between ability groups also varies. For instance, for children from middle income families, input is often received from parents, before commencing school. However, in lower income families, there have often been fewer home experiences with mathematics, apparently leading to a disparity in number skills on entry to school (Jordan et al., 2007). The gap between children from different backgrounds was not found to reduce with instruction. However, in a Chinese study, Zhou (2006) found that Kindergarten teaching was crucial in developing children's number concept (defined as cardinality, written number symbols and addition and subtraction operation). They found that the advantage that children with better educated mothers had on entry to Kindergarten compared to children from working families reduced after a year of teaching.

How numbers are used in the developmental environment impacts on children's learning. Using the words "1, 2, 3...." during structured activity or play demonstrates a positive relationship with early numeracy skills while simply reciting numbers had a negative relationship (Blevins-Knabe & Musun-Miller, 1996). Working closely on number competencies with high risk children on entry to formal education has been shown to improve performance in future years (Griffin, Case & Siegler, 1994).

The current study

This study aims to explore whether symbolic number identification development has a discernible pathway that all pupils progress through. To our knowledge, no such attempt has been made to examine this previously, and we hypothesise that number identification ability may represent a measurable beginning of progress in mathematical concept development. Current research suggests that number identification may be learned progressively, starting with the most common single digits before expanding into multi-digit recognition with increasing experience. We expect that a significant leap may occur when children move from single to two-digit numbers (possibly reflecting the beginning of informal place value knowledge). Subsequently, there will be a leap to three-digit numbers, albeit less substantial as place value principles begin to be consolidated. Higher numbers are expected to follow more easily but still in discrete jumps. As the number of digits increases, necessary vocabulary is acquired, and consolidation of place value continues. One could predict that similar leaps in difficulty would be expected as progression expanded further into centuries, millennia etc. Regardless, contemporary evidence thus far suggests a learning pathway underlying this important numerical component. Its predictive power in young children perhaps suggests that progression along this pathway is the gateway to greater mathematical understanding and may thus be supportive of previous works (Purpura et al., 2013).

Using Rasch modelling, we demonstrate, using a large representative sample of English and Scottish children entering formal education, that such a pathway is unidimensional and describes, the order and stages in which children begin to identify numbers. Evidence suggests that factors such as poverty, culture and language may impact upon a child's progress in mathematics. However, no *a priori* reason exists to expect that this should impact upon the order in which they learn to identify number symbols, only their level of progress. We test the assumption of order invariance across social and demographic groups using differential item functioning (DIF). Furthermore, by examining data from the beginning and the end of the school year we demonstrate the measure remains invariant after instruction, and that children do not regress on number identification performance tasks on reassessment. Finally, we demonstrate via simulation that our analysis is not tautological; i.e. the order we present number symbols in does not determine estimated difficulty.

Method

Sample

11,185 children who started school in England and Scotland in the academic years 2011 and 2012 were analysed. Data came from schools participating in the PIPS (Performance Indicators in Primary Schools) monitoring system run by the Centre for Evaluation and Monitoring (CEM) at Durham University, UK (see www.cem.org and Tymms (1999) for information). This monitoring system provides assessments which schools administer and upload data to CEM for processing. Norm-referenced scores were returned to schools to inform teachers' practice and for self-evaluation purposes. Participation was voluntary, and schools paid an annual registration fee. All schools confirmed that they had provided

sufficient information to parents/guardians about the PIPS system, including the right to opt out. Schools were also informed that anonymised data could be used for research purposes.

For this study, data sets which were representative of pupils in the two countries were generated from the full PIPS dataset by ensuring (via proportional random sampling) equal samples of pupils were found in socio-economic deciles separately for England and Scotland. For details of how representativeness was established see Tymms, Merrell, Hawker and Nicholson (2014).

Pupils' ages were recorded in months and days at the time of assessment. Due to differences between Scottish and English education systems, the average starting age for Scottish children is about half a year later. Only pupils who completed both assessments, start (SOR) and end of reception (EOR), and for which item-level data were available, were included in the analysis; see Table 1. The final sample consisted of 9439 pupils with both SOR and EOR data (49.6% male). 61% of data came from the Scottish sample.

Table 1 about here

Free school meal entitlement (FSM) and having English as an additional language (EAL) data were available. Approximately 7% of pupils in the sample were recorded as EAL, whilst 10.05% were FSM eligible.

Measure

Children's ability to identify numbers was assessed at the SOR using the PIPS On-Entry baseline assessment from the PIPS monitoring system. The assessment includes sections assessing language and mathematics development. It is computer-delivered. Teachers assess one child at a time, presenting questions verbally using recorded sound files. For the Number Recognition section, the child sees a Hindu-Arabic numeral on-screen and is asked 'What is

this number?' to which they respond verbally. Teachers record pupil answers on-screen. Numbers are presented in an approximate order of increasing difficulty, beginning with the number '4' and proceeding through single digits as follows: 1, 3, 2, 5, 7, 6, 9, 8, 0. They then progressed through three randomly selected teens, three randomly selected two-digit and five randomly selected three-digit numbers. Children could be presented up to a maximum of 21 items. Internal reliability (Cronbach's alpha) of the number identification section is 0.93 (Tymms et al., 2012). When the child answers three consecutive items incorrectly, or four wrong in total, the section is terminated. There were no time restrictions on any single item or the assessment section.

Data analysis

Rasch measurement was used to explore the psychometric properties of the assessment. Pupil and item abilities for Number Recognition were estimated and the fit of the data to the Rasch model was investigated at both the SOY and EOY. DIF was examined between SOR and EOR assessments, gender of pupils, age of pupils and geographical location. Changes in person ability between the SOR and EOR assessments were analysed. Data analysis was conducted using WINSTEPS 3.90 (Linacre, 2015).

Analysis assumptions, limitations and caveats

Items with fewer than 20 responses were excluded. Pupils were excluded if they were outside a predetermined age range at the time of administering the SOR assessment (in England, pupils were if their aged between 4.0 and 5.0 years; in Scotland, 4.5 to 5.5 years), or if their age was missing. Some categories were collapsed to ensure sufficient data for analysis. Of the possible 1000 items for numbers 0 to 999, the following groupings were applied:

- All single digit values were included from 0 to 9;

- All two-digit values were included from 10 to 99;
- Multiples of 100 were grouped into one category from 100 to 900 and are represented as X00;
- All three-digit values were grouped according to their first digit from 100s to 900s and represented as 1XX, 2XX, 3XX etc.;
- Three-digit items were analysed according to their position in the assessment. If the item was the first in the group of 3-digit numbers, it was followed by _1 to show this position. Similarly, items in position 2 were followed by _2, items in position 3 by _3 etc.

Analysis strategy

The dichotomous Rasch model (Rasch, 1960, Bond and Fox, 2015) was applied separately to the SOR and EOR data sets. This was because children were assessed twice, and the model assumes that the cases are independent. Although the one-parameter Rasch model is, technically, the simplest of Item Response Theory (IRT) models, it is viewed by many (Bond & Fox, 2015; Wright 1997) as distinctive corresponding to fundamental measurement. From this perspective, the analysis plan is to see if the data fit the model; not to fit the model to the data. For example, suppose a two-parameter model fit the data well and was accepted by the researchers. The two-parameter model uses one parameter for the item/person difficulty/ability and another for the discrimination of items. This would imply that some items varied in difficulty according to the ability of students, contradicting the paper's hypothesis that there is a single scale. We sought to see if the data fit the Rasch model to challenge the hypothesis directly. One-parameter Rasch measurement locates person abilities and item difficulties on the same equal interval scale, presented in an item-person map. If items fit the model well, it provides evidence that a single latent trait has been measured. The

fit of the data to the model was investigated with reference to infit and outfit mean square (MNSQ) statistics.

To further assess unidimensionality, a principal components analysis (PCA) of residuals was undertaken for the SOR and EOR data sets. Whilst the data are not expected to match the Rasch model perfectly, the identification of separate residual components would suggest the assessment was measuring multiple dimensions (Wright, 2000). Having extracted the explained variation in the latent measures due to the underlying trait, the remaining residuals are unexplained model "noise". If additional measures to the main dimension emerge, they will be present in this "noise".

Person separation reliabilities are reported for both the SOR and EOR data. These correspond to Cronbach's alpha as a measure of internal consistency.

Results

Summary statistics from SOR and EOR assessments are reported in Table 2.

Table 2 about here

Person and item reliabilities were high suggesting item difficulties were fixed at zero for both SOR and EOR assessments. At the SOR assessment, the mean of the person abilities was 8.42 logits lower than the mean of the item difficulties. By the EOR, mean person abilities were 1.28 logits lower than the mean of the item difficulties.

As items are presented sequentially in the presence of stopping rules, earlier responses will impact to some extent on subsequent responses. Multiple tests for local item dependence (LID) were performed. There was no evidence of the presence of substantive LID (i.e. sufficient LID to affect the conclusions) and no evidence of it influencing the pattern of item difficulties (see Technical Appendix for further details).

Extreme pupils and items were excluded from Table 2. At the SOR, 13 pupils (0.1%) gained such high scores that they couldn't be placed on the scale, and 600 pupils (6.4%) did not answer any items correctly. Eleven items could not be placed on the scale due to extreme scoring. Three items had no responses because they were not randomly selected during the assessment. In the EOR assessment, 996 pupils (10.6%) had extreme maximum scores (all 21 items identified correctly). No items were found to be extreme at the EOR, although two lacked responses. The person-item maps for the SOR and SOR assessments are shown in Figures 1 and 2.

[Insert Figures 1 and 2 here]

Figures 1 and 2 show item difficulties (right) against person abilities (left). Easier items are towards the bottom of the scale. There are clear jumps in difficulty between numbers 1 to 5 and 6 to 9 (SOR = 2.20 logits/EOR = 2.98 logits), with a second jump between the single digits and teens (SOR = 4.14 logits/EOR= 4.20 logits), a small jump between the teens and other two-digit numbers, (SOR = 1.34 logits/EOR = 2.98 logits) followed by a further jump between two and three-digit numbers (SOR= 4.36 logits/EOR = 6.13 logits). While actual logit sizes of the difficulty gaps differ between the SOR and EOR, the overall pattern across Figures 1 and 2 are indicative of the hypothesised order of difficulty.

Fit statistics for the SOR assessment showed the data fit the model well. The average score was 8.5 (SD = .2). Average infit and outfit MNSQs were .82 and .80 respectively. Model fit statistics for the EOR assessment also showed a good fit between data and the Rasch model. Average infit and outfit MNSQs were .72 and .56 respectively.

The Principal components analysis (PCA) is presented in Table 3. At the SOR, 74.2% of the variance is explained by the measures. The largest secondary dimension exhibits an

Eigenvalue of 4.0 and explains just 0.8% of the variance. At the EOR, the Rasch dimension explains over 80% of the variance.

Table 3 about here

Evidence from model-fit analysis and PCA suggests that the items for number identification form a single scale. Items demonstrated acceptable infit and correlation values, as well as showing adequate discrimination and expected asymptote values in both SOR and EOR assessments. Table 4 illustrates those digits that showed some signs of problematic fit, in these cases, having outfit MSQ values greater than 2.5.

Table 4 about here

Many of these outfit values are high. Interestingly, these digits are all less than 20. Correlation and infit values show that these items accurately target pupils with abilities close to the ability of the items. Therefore, it appears that there were instances of high-ability pupils occasionally making a mistake on an early item (perhaps expecting something more complicated than identifying the number 1, 2 etc.). As the stopping rules allow pupils to continue even if they have made one or two early mistakes, the impact of such errors on a logit scale of this range (28 logits) will likely have substantive impact on outfit statistics (but not other fit statistics) for very easy items. We suggest that these outfit statistics do not demonstrate substantive problems with the measure in the context of this assessment.

Invariance testing

Items were tested to see if difficulties varied across sub-groups (FSM, EAL and geographical location). We were also interested to see if item difficulties remained constant between SOR and EOR assessments, after pupils have been exposed to teaching.

DIF was investigated using the Mantel-Haenszel statistic. DIF was deemed present in an item if there was both; (1) substantive differences in item difficulty between different groups of more than 0.64 logits and (2) the Mantel-Haenszel statistic was significant at $p < .05$ (Linacre, 2011). SOR and EOR data were analysed separately and DIF was checked for EAL, FSM and geographical location. In no instance was DIF found.

As well as the DIF analysis above, in order to compare item difficulties between SOR and EOR data, they were analysed separately, and item difficulty estimates were compared. If item difficulties from separate analyses are identical and plotted against one another, they will fall on a straight line (y=x), with difficulties centred on 0 logits. To determine if substantive differences between the subsamples is statistically significant, a 'quality envelope' is added to the plot. These are established via standard errors of measurement about each data point (95% confidence bands). An item falling outside the quality envelope suggests a significant change in the item's performance across the two subsamples. There is a reasonable expectation, as 95% confidence limits are used, that 5% of items (approximately 7 items in this assessment) would lie outside the quality envelope due to Type I errors.

Estimates of item difficulty locations at the SOR and EOR are plotted in Figure 3, with the quality envelope indicating 95% confidence limits. SOR and EOR item difficulties were positively correlated ($r = 0.99$, $p<0.001$). The number of items which fall outside the quality envelope is greater than 7 implying there were some statistically significant differences between the SOR and EOR. Some single digit numbers were slightly more difficult at the EOR (for example, 4 became slightly more difficult) and some two-digit

numbers were slightly easier at the EOR (for example, 11). But with such a very high correlation between the measures and such a large dataset generating such narrow confidence intervals, we conclude that, substantively, the difficulties remained almost constant from SOR to the EOR.

Figure 3 about here

Pre-test/post-test changes in person ability

Progress of pupils along the measured variable was investigated. For young pupils following a developmental pathway, we would not expect performance on subsequent administrations of the assessment to regress without exceptional circumstances. In other words, the point reached on the scale is sufficient evidence to indicate that pupils will at least get the same score on the second assessment occasion as on the first.

To investigate changes in pupil performance across the two assessment occasions, data from each pupil from both the SOR and EOR were analysed within a single model. Measures were constructed on all observations simultaneously (Wright, 2003). Table 5 shows a cross tabulation produced to illustrate the number of pupils making progress from early single digits (1, 2, 3, 4, 5) through to three digits.

Table 5 about here

Almost all pupils' progress along the ability scale. Just 168 pupils out of 9439 showed lower ability estimates at the EOR compared with the SOR. Table 4 shows that very few pupils regress to a lower number category (i.e. from teens to single digits). Table 6 shows the proportion of pupils who regressed at the EOR.

Table 6 about here

A small number of children with special needs, who were ill or had unusual home arrangements can be expected to make no progress, or to regress. This appears constant across the groups in Table 6.

Is this finding tautological?

It might be thought that the order of difficulty for number identification that has emerged here is tautological, i.e. by presenting single digits first, they are likely to be answered more frequently and thus appear easier. To address this, we conducted simulations based on the pupils in the original data set. We simulated scores for 300001 pupils ranging along an ability scale from -15 to 15 logits (equally spaced at 0.001 logit intervals) and calculated the probability that a pupil gets a question correct based on original item difficulties. We then simulated a complete dataset of responses for each child and each question. The stopping rule was retrospectively applied, and missing responses were treated as wrong. The two data sets (with and without the stopping rules) were then compared to the original data set. Figure 4 illustrates the original difficulties plotted against the estimated difficulties.

Figure 4 about here

As shown in Figure 4, changes to item difficulties are very small. While some are statistically significantly different, ($p<.05$), they are not substantively different. While it is possible that some items could change position slightly, these changes would fall within bands, not between them (i.e. 4 and 5 may swap position but not 4 and 14 or 24). We are thus confident that our hypothesis regarding the difficulty stages of numbers is supported and have that individual number orders have been reflected with precision. For full details regarding the simulation procedure, see the Technical Appendix.

Discussion

This study had three aims: to demonstrate a clear pathway through the learning of number symbols and, if present, to demonstrate that this pathway was both invariant and can be used as a progress measure over time. We address each of these in turn.

Results from this large, nationally representative data set demonstrated the stages that all pupils progress through as they begin to learn and consolidate numerals. Figures 1 and 2 show a distinct series of stages through number identification. Young children begin with numbers 1 to 5 before making their first leap to numbers 6 to 9. The next stage is achieved, and pupils move on to numbers in the teens. Children then access increasingly difficult two-digit figures before finally accessing three-digit numbers. Each stage is measurably more difficult than the previous. The fact that this pathway is unidimensional allows us to conclude that there may be a universal pathway to the identification of number symbols for young children in England and Scotland. This order of progression complements existing literature on the order of which children learn numbers (Lipton & Spelke, 2005; Mix et al., 2014; Wynn, 1992).

Furthermore, the data was invariant across gender, FSM eligibility, EAL groups and countries. This pathway is therefore not unique to particular groups and is unlikely a result of sampling. This is an important finding and suggests children are learning number symbols in the same order, through the same difficulty stages independently of key social, cultural and demographic factors. Item difficulty also remains largely constant (with non-substantive deviations) between the start and end of the first year at school. This is important because it shows that teaching does not (without further contrary evidence) appear to influence the measure and thus the order of acquisition.

Across the two time periods (start of year to end of year), almost every child demonstrated large, measurable progress. Children who take this assessment at the beginning

of the year improve their performance at the end of the year. This is again supportive of existing literature that suggests that children make much progress in this domain during the first year of formal education (Chard et al., 2005).

These combined findings appear to be consistent with progression through a single pathway to number identification. Development is demonstrable within and between the distinct identified stages, i.e. progression is seen through single digits and then another distinct phase appears to be tens, then hundreds and so on. Although we currently cannot specify how the learning happens within stages, the empirical identification of such distinct steps is a major contribution to the literature. Recall also from earlier discussion that number identification is predictive of future attainment in young children as well as demonstrating high reliability, concurrent validity and growth during the early years (Chard et al., 2005, Clarke & Shinn, 2004; Jordan; 2007; Lembke & Foegen, 2009; Tymms, 1999, Tymms et al., 2012). Number identification is a strong correlate of informal 'number sense' concepts, (verbal and one-to-one counting, quantity discrimination, cardinality and subitizing), as well as formal mathematical skills such as addition and subtraction problems (Clarke & Shinn, 2004; LaFevre et al., 2010; Purpura et al., 2013). Number identification also features prominently in early education screening batteries (Jordan, Glutting, Ramineni & Watkins, 2010). The relationship between number identification and later formal skills and strategies (such as count-up/back, derived and known facts) is less well understood so far as the authors can tell. While number sense batteries in longitudinal studies (see Jordan et al., 2010; LaFevre et al., 2010 for examples) use number identification tasks, these are often part of latent variables where its unique contribution is often unexamined. A one-year longitudinal study by Gobel, Watson, Lervag and Hulme (2014) on six-year-old UK pupils suggested that number identification uniquely predicted arithmetic proficiency over time. Similar results were found in Finnish children (Zhang, Räsänen, Koponen, Aunola, Lerkkanen & Nurmi,

2017). However, further work is necessary to examine the impact of number identification on advanced mathematical strategies in later childhood. Despite these gaps in the current literature, the significance of number identification as a teaching and learning tool cannot be understated. It may be that number identification can be viewed as a practical, universal progress measure that has major implications for both empirical research and practical teaching strategies.

We provide evidence that these effects are not artefacts of the assessment stopping rules or due to item presentation order. If the presentation order was not the order of difficulty, pupils would unexpectedly get apparently hard items right and the Rasch model would demonstrate this. If the presentation order was the driving force in this study, we would expect to see the this reflected in the difficulties of the first ten digits (recall that they were presented in the order 4, 1, 3, 2, 5, 7, 6, 9, 8, 0). Our results suggest that the relative order of difficulty was in fact 1, 3, 4, 2, 5, 7, 0, 6, 8, 9 at the start of the year and 1, 2, 3, 4, 5, 7, 0, 6, 8, 9 at the end of the year. Simulation data confirms our items positioning along the logit scale. While some numbers (particularly numbers such as 100) appear easier than expected, it could be that these are just easily recognisable numbers (which are perhaps common in our developing environments; Byrge et al., 2014; Dehaene & Mehler, 1992; Mix et al., 2014) and does not necessarily suggest leaps in general numerical understanding. The presented pathway is thus likely not artefactual of the measure and represents the relative order of difficulty of identifying Hindu-Arabic number symbols.

Limitations and future directions

The prospect of a universal progress measure is exciting but much work must be done to verify this study's conclusions. Importantly, the sample analysed is currently restricted to pupils in England and Scotland between ages of 4 – 6.5 years. Generalising beyond these

ages, nations, and whether this pattern holds for number symbols in the number ranges in the thousands and beyond, will require a continuation of this work.

This study focused exclusively on the Hindu-Arabic number system. Additional work should examine other numerical symbol systems (e.g. Chinese, Bengali etc.). If the same pattern exists in other numerical systems, it may tell us more about how children mentally represent and map number symbols. Learner's language is also crucial. While we exhibit no evidence of DIF between English learners versus EAL speakers (possibly due to a small, heterogeneous group covering all non-English languages), we reviewed evidence that suggests specific language groups may differ from each other (Pixner et al., 2011; Seron et al., 1994). How this impacts upon early symbol learning remains unexplored and repeating this assessment on representative samples from other specific language groups may be illuminating.

Methodologically, our assessment introduced the potential for tautological findings, and may have increased the likelihood of LID, which could subsequently impact upon estimates. While analysis suggest these are not serious issues in this data set, confirmation of these findings via different methods would serve to strengthen this study's findings.

Finally, we reiterate that the findings do not extend our knowledge on how children learn numerosity. Number identifications one of many requisite skills that may act as a gateway into the effective learning of formal mathematics (Purpura et al., 2014). This study provides evidence to suggest that there is a specific pathway through number symbol identification. How children progress to map numerosity to corresponding symbols remains an important question.

Implications and conclusions

We have presented evidence for the order in which children learn to identify numbers being a universal progress measure from which conclusions can be drawn regarding children's' mathematical development. This has implications for early years teaching and learning. The assessment of pupils' ability in number recognition on a more frequent basis throughout their first year in education may be one approach to identifying the progress of pupils within and between stages. From this information, teachers can tailor activities appropriate to the stage of each child's developmental level. While this may be considered common knowledge (at least now an empirical verification thereof) and practice, this study does offer novel insight into how children learn number. The results suggest that children learn to identify the first five digits and will not progress to learning the second half (6 to 9) until they have consolidated their learning of the former. Once confident with the identification of digits within an identified stage, they need to engage in activities to practice and become familiar with digits in the next stage. These discrete jumps are not necessarily intuitive, and it may be informative for practitioners to understand that the difficulty of moving from 5 to 6 is large, relative to then moving from 36 to 96, which are approximately equivalent (See Figures 1 & 2).

In England, the 'Development Matters in the Early Years Foundation Stage' publication is non-statutory guidance material to support practitioners in implementing the statutory requirements of the Early Years Foundation Stage (Early Education, 2012). This guidance (supported by the Department for Education) includes a description of typical development for children of different ages from birth to 60 months. This guidance to support children's development of number recognition is imprecise as indeed is the Statutory Framework for the Early Years Foundation Stage (Department for Education, 2014). In Scotland the Curriculum for Excellence Experiences and Outcomes guidance is similarly imprecise in setting out a developmental progression for number recognition

(www.educationscotland.gov.uk). These documents could benefit from including more detailed information on the basis of the findings of this study. Our findings have clear implications for policy-makers who could use them to inform curricula for the early years. If, as some have suggested, number identification is a gateway concept into further mathematics, the impetus to ensure children learn and consolidate their knowledge in this area is critical.

References

Baroody, A.J., & Wilkins, J.L.M. (1999). The development of informal counting, number and arithmetic skills and concepts. In J.V. Copley (Ed.), *Mathematics in the early years* (pp. 48-65). Washington, DC: National Association for the Education of Young Children.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78, 551–572.

Berch, D.B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*; 38, 333–339. doi.org/10.1177/00222194050380040901

Bertelli, R., Joanni, E., & Martlew, M. (1998). Relationship between children's counting ability and their ability to reason about number. *European Journal of Psychology of Education*, 8, 371-384. doi.org/10.1007/BF03172951

Blevins-Knabe, B., & Musun-Miller, L. (1996). Number use at home by children and their parents and its relationship to early mathematical performance. *Early Development and Parenting*, 5, 35–45. doi.org/10.1002/(SICI)1099-0917(199603)5:1%3C35::AID-EDP113%3E3.0.CO;2-0

Bond, T.G., & Fox, C.M. (2015) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Third Edition). New York: Routledge.

Butterworth, B., & Reigosa, V. (2007). Information processing deficits in dyscalculia. In D.B. Berch & M.M.M Mazzocco (Eds.). *Why is math so hard for some children? The nature and origins of mathematical learning difficulties* (pp. 65-81). Baltimore, MD:Brookes.

Byrge, L., Smith, L.B., & Mix, K.S. (2014). Beginnings of place value: How pre-schoolers write three-digit numbers. *Child Development*, 85, 437-443. doi:10.1111/cdev.12162

Carey, S. (2004). Bootstrapping and the origins of concepts. *Daedalus,* 133, 59-68.

Chard, D.C., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, 30, 3-14. doi.org/10.1177/073724770503000202

Clarke, B., & Shinn, M.R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234-248.

Dehaene S, Mehler J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, 43, 1–29. Doi: 10.1016/0010-0277(92)90030-L [PubMed: 1591901]

Department for Education (2014). *Statutory framework for the early years foundation stage*. www.gov.uk/government/publications.

Early Education (2012). *Development Matters in the Early Years Foundation Stage (EYFS)*. Pub. Early Education: London. www.early-education.org.uk ISBN 978-0-904-187-57-1

Fuson, K.C., (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit additions, subtraction and place value. *Cognition and Instruction*, 7, 343-403.

Gelman, R., & Gallistel, R. C. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gersten, R., Jordan, N.C., & Flojo, J.R. (2005). Early Identification and Interventions for Students With Mathematics Difficulties. *Journal of Learning Disabilities*, 38, 293-304. doi.org/10.1177/00222194050380040301

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability* (3rd ed.). Austin, TX: Pro-Ed.

Gobel, S.M., Watons, S.E., Lervag, A., & Hulme, C. (2014). Children's arithmetic development: It is number knowledge, not the approximate number sense, that counts. *Psychological Science*, 25, 789-798. doi: 10.1177/0956797613516471

Griffin, S., Case, R., & Siegler, R.S. (1994). Classroom lessons: Integrating cognitive theory and classroom practice. In: McGilly, K., (Ed). *Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure*. Cambridge, MA: MIT Press; 1994. p.25-50.

Jordan, N.C., Glutting, J., Ramineni, C., & Watkins, M.W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, 39, 181-195.

Jordan, N.C., Kaplan, D., Locuniak, M.N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, 22, 36-46. doi.org/10.1111/j.1540-5826.2007.00229.x

Jordan, N.C., Kaplan, D., Ramineni, C., & Locuniak, M.N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, 45, 850-867. doi.org/10.1037/a0014939

Krajewski, K., & Schneider, W. (2009). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: Findings from a 3-ytear, longitudinal study. *Journal of Experimental Child Psychology*, 103, 516-531

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition,* 105(2). doi:10.1016/j.cognition.2006.10.005.

LeFevre, J.A., Fast, L., Smith-Chant, B.L., Skwarchuk, S.L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to Mathematics: Longitudinal Predictors of Performance. *Child Development*, 81, 1753-1767. doi.org/10.1111/j.1467-8624.2010.01508.x

Lembke, E., & Foegen, A. (2009). Identifying Early Numeracy Indicators for Kindergarten

and First-Grade Students. *Learning Disabilities Research & Practice,* 24, 12–20

Linacre, M. (2015). *Winsteps Rasch Measurement* 3.90. www.winsteps.com

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large number discrimination in human infants. *Psychological Science*, 14, 396 – 401. doi.org/10.1111/1467-9280.01453

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, 5, 271 – 290. doi.org/10.1207/s15327078in0503_2

Lipton, J. S., & Spelke, E. S. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child Development*, 76, 978-988. doi.org/10.1111/j.1467-8624.2005.00891.x

Malofeeva, E., Day, J., Saco, X., Young, L., & Ciancio, D. (2004). Construction and evaluation of a number sense test with Head Start children. *Journal of Educational Psychology*, 96, 648–659. doi.org/10.1037/0022-0663.96.4.648

Martin, R. B., Cirino, P. T., Sharp, C., & Barnes, M. (2014). Number and counting skills in kindergarten as predictors of grade 1 mathematical skills. *Learning and Individual Differences*, 34, 12–23. doi.org/10.1016/j.lindif.2014.05.006

Mazzocco, M.M. & Thompson, R.E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice*, 20, 142–155. doi.org/10.1111/j.1540-5826.2005.00129.x

Merkley, R., & Ansari, D. (2016). Why numerical symbols count in the development of mathematical skills: evidence from brain and behaviour. *Current Opinion in Behavioral Science,* 10, 14-20.

Miller, K.F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States: Cross-cultural insights into the nature and the course of preschool mathematical development. In Campbell, J.I.D. (Ed), *Handbook of Mathematical Cognition*, pp 163-178. New York: Psychology Press.

Miller, K.F., Smith, C.M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross national differences in mathematical competence. *Psychological Science*, 6, 56-60. doi.org/10.1111/j.1467-9280.1995.tb00305.x

Mix, K.S. (2009). How Spencer made number: First uses of the number words. *Journal of Experimental Child Psychology*, 102, 427-444. doi.org/10.1016/j.jecp.2008.11.003

Mix, K.S., Huttenlocher, J., & Levine, S.C. (2002). *Quantitative development in infancy and early childhood.* New York: Oxford University Press

Mix, K.S., Prather, R.W., Smith, L.B. & Stockton, J.D. (2014). Young children's interpretation of multidigit number names: From emerging competence to mastery. *Child Development*, 85, 1306-1319.

Mussolin, C., Nys, J., Content, A., & Leybaert, J. (2014). Symbolic number abilities predict later approximate number system acuity in preschool children. *PLOS one, 9, e91839*

National Mathematics Advisory Panel. (2008). *Foundations for success: Final report of the National Mathematics Advisory Panel*. Washington, DC: United States Department of Education.

Pixner, S., Zuber, J., Hermanova, V., Kaufmann, L., Nuerk, H.-C., & Moeller, K. (2011). One language, two number systems and many problems: Numerical cognition in the Czech language. *Research in Developmental Disabilities*, 32, 2683-2689.

Purpura, D.J., Baroody, A.J., & Lonigan, C.J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology,* 105, 453-464.

Purpura, D.J. & Ganley, C.M. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology*, 122, 104-121. doi.org/10.1016/j.jecp.2013.12.009

Purpura, D.J. & Reid, E.E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, 36, 259–268. doi.org/10.1016/j.ecresq.2015.12.020

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rouselle, L., & Noel, M. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs. non-symbolic number magnitude. *Cognition*, 102, 361-395.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: learning trajectories for young children*. New York, NY: Routledge.

Seron, X., & Fayol, M. (1994). Number transcoding in children: A functional analysis. *British Journal of Developmental Psychology*, 12, 281-300.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, 75, 428–444. doi.org/10.1111/j.1467-8624.2004.00684.x

Song, M. J., & Ginsburg, H. P. (1987). The development of informal and formal mathematical thinking in Korean and U.S. children. *Child Development*, 58, 1286–1296.

Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading*, 22, 27-36. doi.org/10.1111/1467-9817.00066

Tymms, P., Merrell, C., Hawker, D., & Nicholson, F. (2014). *Performance indicators in Primary Schools: A comparison of performance on entry to school and the progress made in the first year in England and four other jurisdictions*. Department for Education, London https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/318052/RR344_-_Performance_Indicators_in_Primary_Schools.pdf

Tymms, P., Merrell, C., Henderson, B., Albone, S. & Jones, P. (2012). Learning Difficulties in the Primary School Years: Predictability from On-Entry Baseline Assessment. *Online Educational Research Journal,* June 2012.

Wagnar, S.H., & Walters, J. (1982). A longitudinal analysis of early number concepts. In G. Foreman (Ed.) *Action and though: From sensorimotor schemes to symbolic operations* (pp. 137-161) New York: Academic.

Wright B. D. (1997) Fundamental Measurement *Rasch Measurement Transactions,* 11:2 p. 558.

Wright, B.D. (2000). Conventional factor analysis vs. Rasch residual factor analysis. *Rasch Measurement Transactions*, 14, 753.

Wright, B.D. (2003). Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test. *Rasch Measurement Transactions*, 17, 905-906.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, 24, 220 – 251. doi.org/10.1016/0010-0285(92)90008-P

Zhang, X., Räsänen, P., Koponen, T., Aunola, K., Lerkkanen, M., & Nurmi, J. (2017). Knowing, applying, and reasoning about arithmetic: Roles of domain-general and numerical skills in multiple domains of arithmetic learning. *Developmental Psychology*, 53, 2304-2318.

Zhou, X. (2006). Children's representation of written number symbols. *International Journal of Early Childhood Education*, 12, 5-21.

Table 1: Sample characteristics (age)

		All		Male		Female	
		N	Age *(SD)* (yrs)	*N*	Age (*SD*) (yrs)	*N*	Age (*SD*) (yrs)
England	SOR	3653	4.52 *(0.28)*	1896	4.52 *(0.27)*	1757	4.52 *(0.28)*
	EOR		5.28 *(0.28)*		5.28 *(0.27)*		5.28 *(0.28)*
Scotland	SOR	5786	5.04 *(0.28)*	2788	5.04 *(0.27)*	2998	5.04 *(0.27)*
	EOR		5.78 *(0.28)*		5.78 *(0.27)*		5.77 *(0.27)*

Table 2: Summary statistics

	SOR		EOR	
	Pupils	Items	Pupils	Items
Measure (logits)				
Mean *(SD)*	-8.42 *(5.04)*	0.00 *(5.31)*	-1.28 *(4.91)*	0.00 *(5.20)*
Range	-17.48 to 9.05	-17.06 to 8.78	-17.21 to 7.91	-16.81 to 7.32
Person Reliability	0.90	0.99	0.82	1.00
Separation	2.98	8.93	2.13	22.39

Table 3: Standardised residual variance at start and end of reception

	SOR	EOR
Raw variance explained by the measures		
Eigen value (%)	388.9 (74.2%)	617.3 (80.8%)
Raw variance explained in the first contrast		
Eigen value (%)	4.0 (0.8%)	1.7 (0.2%)
Raw variance explained in the second contrast		
Eigen value (%)	3.0 (0.6%)	1.6 (0.2%)

Table 4: Items with high outfit values

	SOR				EOR			
Digit	Difficulty	Infit MSQ	Outfit MSQ	Correlation	Difficulty	Infit MSQ	Outfit MSQ	Correlation
4	-16.16	1.26	9.90	0.58	-15.17	1.29	9.90	0.26
1	-17.06	1.22	9.90	0.53	-16.81	1.32	0.64	0.21
3	-16.08	0.90	9.90	0.60	-15.66	0.84	0.28	0.25
2	-15.62	0.79	9.90	0.49	-15.94	0.69	0.02	0.18
7	-12.64	0.84	4.82	0.60	-11.99	0.89	3.16	0.35
6	-11.62	1.05	6.16	0.60	-11.38	1.02	2.33	0.36
9	-10.64	0.99	3.52	0.62	-10.19	0.95	1.45	0.40
8	-11.10	0.83	3.22	0.52	-10.80	0.85	1.83	0.30
0	-11.36	0.99	9.90	0.39	-11.68	1.01	9.90	0.20
11	-7.22	0.90	2.64	0.69	-7.48	0.93	2.78	0.45
12	-5.04	1.11	9.90	0.72	-4.72	1.26	9.90	0.62
13	-3.82	1.00	8.38	0.75	-3.94	1.01	9.90	0.69
14	-5.36	0.85	9.90	0.76	-5.59	1.03	9.90	0.58
15	-3.42	0.84	1.52	0.77	-4.22	0.90	9.90	0.70
16	-5.03	0.82	9.90	0.77	-5.08	0.96	9.90	0.64
17	-4.90	0.74	9.90	0.78	-5.08	0.92	9.90	0.64
18	-4.74	0.81	2.92	0.78	-4.98	0.99	9.90	0.65

Table 5: Cross tabulation of pupil's performance from SOR to EOR

		Start of year pupil performance				
		Single digits (1,2,3,4,5)	Single digits (6,7,8,9,0)	Teens	Two digit	Three digit
End of year pupil performance	Three digit	88	848	2240	880	393
	Two digit	254	1090	1007	101	0
	Teens	714	1021	237	8	0
	Single digits (6,7,8,9,0)	281	83	4	0	0
	Single digits (1,2,3,4,5)	186	4	0	0	0

*Darkness of shading represents progress i.e. darker shades represent greater progress

Table 6: Proportion of pupils who gained lower ability measures across the academic year

	Area		Gender		Age	
	England	Scotland	Male	Female	Young	Old
N	3653	5786	4684	4755	4351	5088
Frequency of lower scores	59	109	84	84	73	93
% of group	1.62	1.88	1.79	1.77	1.68	1.83
% of lower scores	35.12	64.88	50.00	50.00	43.45	55.36

Figure 1: Person-item map at SOR

```
MEASURE    PERSON - MAP - ITEM
             <more>|<rare>
   10              +
                   |
    9              .  +  8xx_2
                   .  |  8xx_4
    8              .  +  2xx_2  3xx_5  4xx_2  6xx_5
                   .  |  2xx_4  2xx_5  7xx_1
    7              .  +  1xx_5  4xx_1  5xx_2  6xx_3  7xx_2
                   .  |  1xx_4  2xx_3  3xx_1  3xx_3  5xx_4  6xx_2  7xx_3  7xx_4  8xx_1  9xx_1  9xx_3
    6              .  +  5xx_1  6xx_1  8xx_3  9xx_2
                   .  |S 3xx_2
    5              .  +  1xx_3
                   .  |  1xx_2  2xx_1  3xx_4  x00_3
    4              .  +  1xx_1
                   .  |
    3            .##  +
                   .  |
    2              .  +  x00_2
                   . T|
    1              .  +
                   .  |  89
    0              .#  +M 31 36 37 51 59 67 72 75 78 86 91 94 95 97
                   .  |  24 26 29 32 34 38 39 40 46 52 53 54 56 57 58 64 68 71 73 74 76 79 80 82 84 87 88 98 x00_1
   -1              .#  +  21 23 27 28 30 35 41 47 48 49 50 55 60 61 62 65 66 70 77 81 83 90 92 93 96
                   .#  |  25 42 44 45 63 69 85
   -2                  +  43
                 .###  |  22 33
   -3          .####  +
                 .## S|  15 20
   -4             .##  +  13
                 .###  |  18
   -5            .###  +  12 16 17
                  .##  |S 14
   -6            .###  +
                  .##  |
   -7                .  +  11
     .#############  |
   -8         .######  +
               .##### M|
   -9                  +
                       |
  -10     .#########  +
                      |T 9
  -11     .#########  +  8
                      |  0 6
  -12         .#####  +
               .####  |  7
  -13                 +
          .######## S|
  -14                 +
               .###  |
  -15                 +  5
               .###  |  2
  -16                 +  3 4
              .####  |
  -17                 +  1
              .####  |
  -18
              <less>|<frequent>
 EACH "#" IS 75: EACH "." IS 1 TO 74
```

Figure 2: Person-item map at EOR

```
MEASURE    PERSON - MAP - ITEM
             <more>|<rare>
    8              +
                .##  |  8xx_1
    7             .  +  6xx_1  7xx_1  9xx_1
                .#   |  3xx_1  4xx_1  5xx_1  5xx_4  6xx_2  7xx_4  9xx_4
    6             .  +  2xx_1  2xx_4  3xx_2  3xx_3  3xx_5  4xx_2  4xx_3  5xx_2  5xx_3  5xx_5  6xx_3  6xx_4  6xx_5  7xx_2  7xx_3  7xx_5  8xx_2  8xx_3  8xx_4  9xx_2  9xx_3  9xx_5
                .#   |  1xx_4  1xx_5  2xx_2  2xx_3  2xx_5  3xx_4  4xx_4  4xx_5  8xx_5  x00_4  x00_5
    5             .  +S 1xx_1  1xx_2  1xx_3
                .#   |
    4             .  +
                  . S|  x00_3
    3             .  +  x00_2
      .############   |
    2       .#######  +
                  .  |  x00_1
    1             .  +
                  .  |
    0             .  +M 31
                .####  |  34 37 57 71
   -1             .##  +  21 23 30 35 36 38 39 49 51 52 53 54 58 59 60 61 74 80 81 86 90 91 96
                   .# M|  24 25 26 27 28 29 32 33 41 45 50 55 56 65 67 68 69 70 72 75 76 77 78 79 84 85 87 88 89 92 93 95 97 98
   -2           .####  +  40 42 43 46 47 48 62 63 64 66 73 82 83 94
                  .   |  22 44
   -3           .####  +
                .####  |
   -4             .#   +  13 15 20
                 .##   |  12
   -5             .#  +S 16 17 18
                  .#   |  14
   -6             . S+
                  .#   |
   -7                  +
                .####  |  11
   -8             .#   +
                  .#   |
   -9                  +
                  .#   |
  -10             .    +  9
                  .    |T
  -11             . T+  8
                  .    |  0 6
  -12             .    +  7
                  .    |
  -13             .    +
                  .    |
  -14                  +
                  .    |
  -15                  +  4 5
                  .    |  3
  -16             .    +  2
                       |
  -17             .    +  1
                       |
  -18             .    +
                <less>|<frequent>
 EACH "#" IS 116: EACH "." IS 1 TO 115
```

Figure 3: Start and end of year item difficulties with 95% confidence intervals

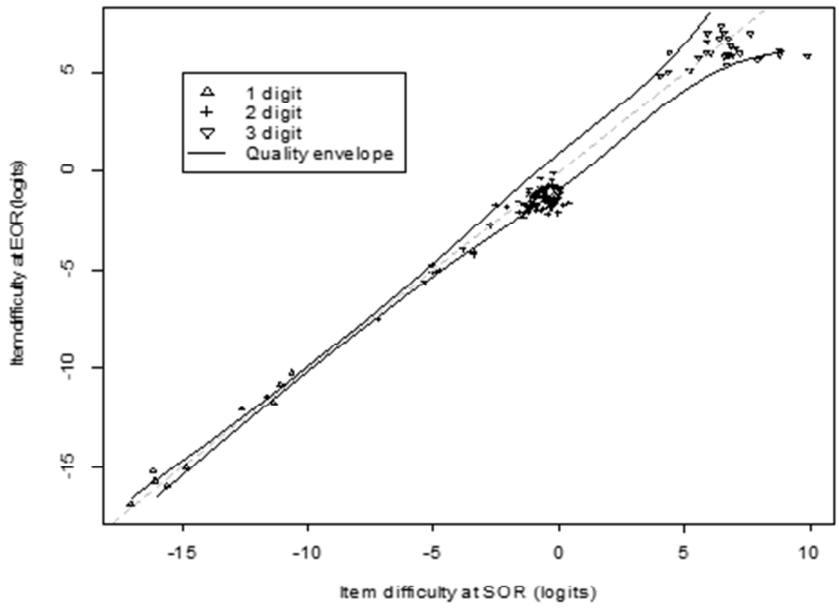
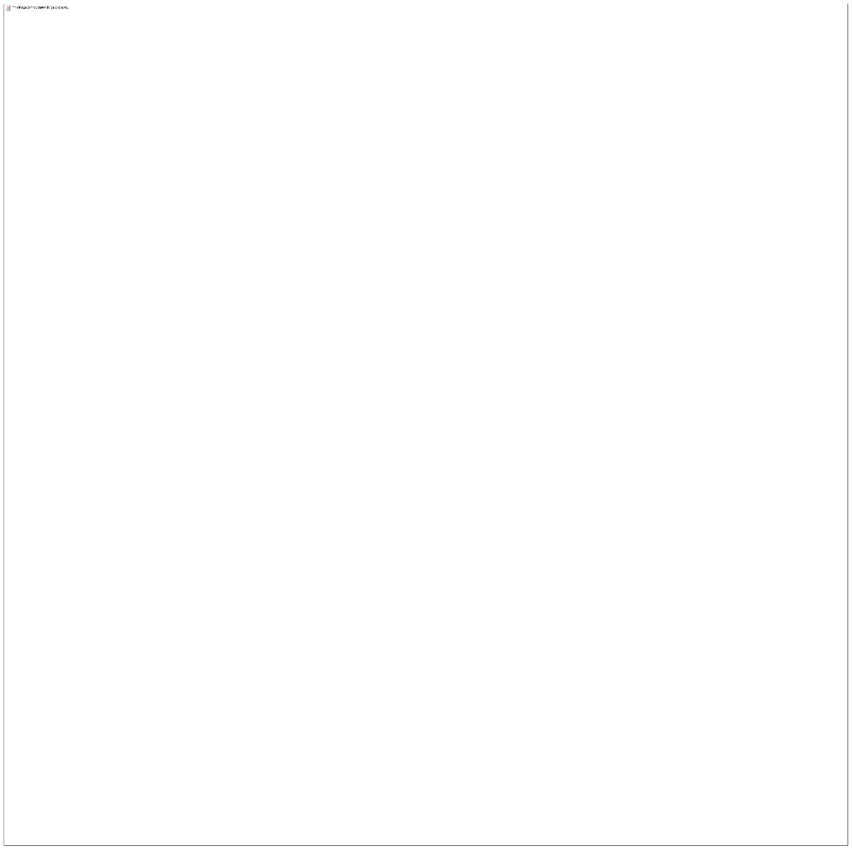


Figure 4: Estimated item difficulty (based on simulation) versus true item difficulty



Technical Appendix

Assumptions of item independence

One of the core assumptions of the Rasch model is that items on a scale should be locally independent of each other i.e. performance on one item should not be dependent on another item for each person. When this assumption is violated, item and person parameters and estimates may be inaccurately estimated.

In this study, given the nature of the task (sequential ordering of digits of increasing size) combined with the inclusion of a stopping rule, it would not be unreasonable to suggest that the assumption of localised item independence may be violated. Evidence of item dependencies can be detected in several ways.

Residuals of observed and expected results

Andrich and Kreiner (2010) detail one such approach and suggest that dependencies can be identified in dichotomous data by calculating the correlation between residuals of observed and expected responses. Larger correlations would signify higher likelihood of dependency.

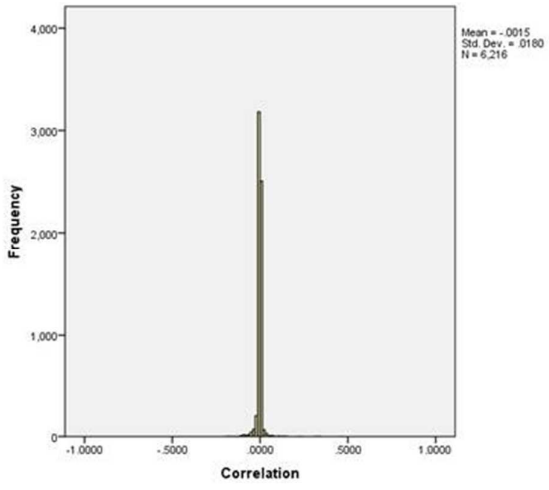


Figure A1: Histogram of residual correlations between observed and expected responses.

Figure A1 shows a histogram of these residual correlations and shows that the mean correlation strength is less than .001 (with a SD of .02). There are no correlations that would be considered moderate or large in terms of their size. This approach does not suggest the presence of major dependencies in this data.

Testlet Comparisons and Q^3

Another form of detection comes from the creation of item parcels, or testlets (Zenisky, Hambleton & Sireci, 2002). Testlets can be created by combining items suspected of dependencies into single items. These can be done either by a predefined structure or randomly. The model is then run using these testlets and the reliability coefficients compared. To examine this, five testlets were created for two models; one with five random testlets and the other based on our hypothesised sequential ability bands; 1-5, 6- 9, 11-19, 20-99 and three-digit items. The item reliabilities of these models were compared with a model of all the dichotomous items. Table A1 shows the item reliabilities of these models. As can be seen, these reliabilities do not differ substantively from each other. While the sequential model appears to be higher, suggesting that there may be some local dependence, the change is very small and provides no evidence that dependencies in this data are large.

Table A1: Reliabilities of Testlet models and Full dichotomous model

	Full set of dichotomous items	Sequential testlets	Random testlets
Persons	0.91	0.93	0.91
Items	0.98	1.00	0.98

Yen (1984) proposed the Q_3 statistic as a measure of dependency detection. Q3 is the correlation of residuals between item pairs once ability has been partialled out. It is calculated for each person and then used to estimate the performance of persons on each item. The residual is calculated as the difference between the observed and the expected results. Q_3 is the correlation of these deviations

across all persons. If there are no local dependencies, Q3 should be approximately -1/(n – 1), where n is number of test items.

Q3 was calculated for the full dichotomous model and then the two testlet models created earlier. As can be seen from Table A2, the observed Q3 values for the full model and random models are very similar. The sequential testlet model shows greater deviations however and suggests that there may be a small amount of dependency in the data set.

Table A2: Q3 statistics for the Full dichotomous model and the testlet models.

	Q_3	Expected	N
Full set	-0.003	-0.007	148
Sequential	-0.02	-0.25	5
Random	-0.22	-0.25	5

Backwards estimation of number difficulties by simulation

Estimating the difficulty of each of the numbers in the PIPS Baseline test is complicated by the application of a stopping rule. Ignoring the structured nature of the missing data could result in inaccurate estimates of item difficulty. Indeed, if we simulate data where all items have the same difficulty (e.g. 0) and then apply the stopping rules, we can clearly see a relationship between the item position and estimated item difficulty (r=0.85). While this correlation here highlights the potential for a tautological finding through measurement artefact, it does not directly help with estimating item difficulties from data with a stopping rule applied.

We can however take this approach in a different direction and attempt to reverse engineer true difficulty estimates by looking at the apparent item difficulties after application of the stopping rule and identifying true item difficulties with associated apparent item difficulties close to those observed. In doing so, we demonstrate the emergence of the reported difficulty bands discussed in

our manuscript even if we cannot be certain of the absolute position of individual numerals within bands.

The process applied is:

1. 3001 pupil abilities are fixed with a uniform spread between -15 and +15 logits. The spread of pupils was intended to ensure some pupils got every item wrong and every item right in each realisation. Having 3001 uniformly spread pupils gives a stable pupil base.
2. Initial difficulties of 0 are assigned for each of the 21 items. Starting with all initial difficulties set to 0 avoids biasing results towards the anticipated outcome.
3. Simulation difficulties are created based on changes to the initial difficulties based on draws from a N(0,2) distribution, where 10% of the items are changed independently in each realisation. Only 10% of the items are changed on any given occurrence because if all items are changed the sample space may become excessively large. The 10% is realised on an item by item basis so typically 2 or 3 item difficulties are changed on each simulation.
4. Full response data is then simulated for the 3001 pupils and 21 simulated difficulties. The stopping rules are then applied to this data for every pupil.
5. The simulated stopped data is then Rasch analysed, treating missing as missing and item difficulties subsequently estimated.
6. These estimated item difficulties are then compared to those observed in the original data.
7. This process is repeated 20 times. The simulated difficulties resulting in the estimated item difficulties which are closest to the observed item difficulties are then identified. We could, instead of taking the best of 20 simulations, take any simulation which is better than the previous one. Very little difference between the final item difficulties would be anticipated. Closeness is measured by the mean squared difference between the observed item difficulties and those estimated from the simulated data after application of the stopping rules.

8. This whole process is then repeated many times (100) with the initial item difficulties replaced with the set giving rise to the closest estimate from the previous realisation on each iteration. 100 full iterations result in a reasonably stable sum of squared errors.

Results

Table A3: 1st realisation based upon 100 iterations

	4	1	3	2	5	7	6	9	8	0	teen	teen	teen	2dig	2dig	2dig	3dig	3dig	3dig	3dig	3dig
Original Observed Difficulties	-9.4	-10.3	-9.2	-8.8	-8.1	-5.8	-4.9	-3.9	-4.3	-4.6	0.3	0.4	0.5	4.4	4.1	4	11.4	11	10.9	11.2	11.1
Difficulties from Simulation	-9.2	-10.6	-10.4	-8.5	-8.8	-5.3	-4.1	-4	-4.5	-5.7	0.6	-0.1	0.4	4.4	4.9	4.9	10.3	11.8	11.2	11.5	11.3
Underlying difficulties	-7.9	-9	-8.7	-7.2	-7.6	-4.4	-3.4	-3.4	-3.8	-4.8	0.6	-0.1	0.3	3.7	4.2	3.9	8	9.5	9	9.3	9

Table A4: 2nd realisation based upon 100 iterations

	4	1	3	2	5	7	6	9	8	0	teen	teen	teen	2dig	2dig	2dig	3dig	3dig	3dig	3dig	3dig
Original Observed Difficulties	-9.4	-10.3	-9.2	-8.8	-8.1	-5.8	-4.9	-3.9	-4.3	-4.6	0.3	0.4	0.5	4.4	4.1	4	11.4	11	10.9	11.2	11.1
Difficulties from Simulation	-8.8	-10.5	-8.1	-10.6	-8.2	-5.2	-6.2	-4	-4.8	-4.1	1.1	0.2	0.6	4.2	3.8	4	11.1	12.3	11.5	11.4	10.4
Underlying difficulties	-6.4	-8	-5.9	-8	-6	-3.4	-4.3	-2.3	-3	-2.5	1.9	0.9	1.4	4.2	3.9	4.3	9.9	10.6	10.1	10	8.9

The underlying difficulty estimates change noticeably between the two set of simulations. However, there is still clear evidence of the discrete ability bands in each simulation.

The third simulation also has 100 simulation steps but starts from the original observed difficulties rather than 0. While the numbers differ slightly from the first two iterations the banding remains clear.

Table A5: Backwards realisation based upon 100 iterations

	4	1	3	2	5	7	6	9	8	0	teen	teen	teen	2dig	2dig	2dig	3dig	3dig	3dig	3dig	3dig
Original Observed Difficulties	-9.4	-10.3	-9.2	-8.8	-8.1	-5.8	-4.9	-3.9	-4.3	-4.6	0.3	0.4	0.5	4.4	4.1	4	11.4	11	10.9	11.2	11.1
Difficulties from Simulation	-9.8	-9.9	-9.2	-9.1	-7.5	-5.1	-5	-3.5	-4.6	-5	-0.1	0.3	1.4	3.4	4.3	3.9	11.4	10.6	11.7	10.9	11
Underlying difficulties	-8.2	-8.1	-7.7	-7.7	-6.1	-4.2	-3.9	-2.7	-3.8	-3.8	0.3	0.4	1.5	3.1	3.9	3.4	9.3	9	9.7	9.3	9

In our forth simulation we run 1000 realisations which gives a sum of squared errors of 1.9 and a correlation of 0.9992 between the original observed difficulties and the difficulties from the simulation.

Table A6: Realisation based upon 1000 iterations

	4	1	3	2	5	7	6	9	8	0	teen	teen	teen	2dig	2dig	2dig	3dig	3dig	3dig	3dig	3dig
Original Observed Difficulties	-9.4	-10.3	-9.2	-8.8	-8.1	-5.8	-4.9	-3.9	-4.3	-4.6	0.3	0.4	0.5	4.4	4.1	4	11.4	11	10.9	11.2	11.1
Difficulties from Simulation	-9.8	-9.9	-9.3	-8.7	-7.7	-5.7	-5.5	-3.8	-4.1	-4.9	0.1	0.5	0.6	4.8	4.2	4.3	11.6	11.1	10.5	11.2	10.4
Underlying difficulties	-7.2	-7.2	-6.7	-6.3	-5.5	-3.7	-3.5	-2.2	-2.4	-3	1.3	1.6	1.9	5.1	4.7	4.7	10.4	10.2	9.9	10.4	9.6

Tracking the difficulty estimates over these 1000 iterations the most striking element is the five bands of questions which are apparent after only 50 iterations. The upper 3 bands show no consistent patterns between items, however, this is entirely consistent with what we would expect since the actual items are a) mixed up within the bands and b) would have less precise estimates because each item would have fewer cases as a result of the stopping rules. The lower 2 bands which relate to the single digit numbers show strikingly consistent internal patterns. There are few instances of overlap between bands after the first iterations.

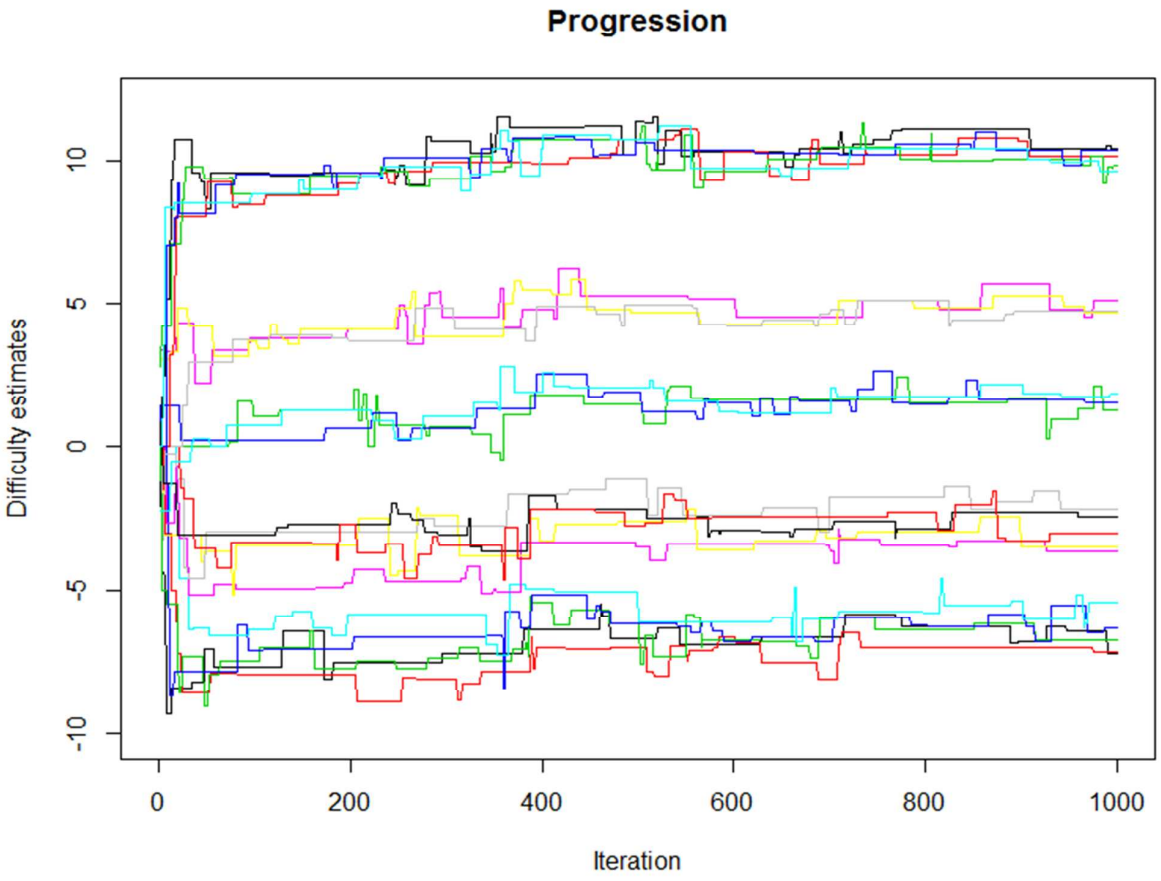


Figure 1: Item difficulties of numbers 1 to 5, 6 to 9, three teens, three 2-digits and five 3-digits over 1000 iterations.

Conclusions

The original observed item difficulties are influenced by the stopping rules, however the underlying difficulties which give rise to these observations retain the same structure with 5 very easy items (numbers 1-5), 5 easy

items (numbers 0 and 6-9), 3 medium items (teens), 3 harder items (2 digit numbers) and 5 very hard items (3 digit items). We therefore assert that while there may be some variation around the specific order of some numbers within each difficulty band, the bands themselves are stable enough to be detected via the employed assessment and that the stopping rules themselves do not manifest this effect as an artefact of the method.

Potential limitations of the method

Alternative solutions are also potentially possible. We may have found one local minima since our starting point and allowed steps maybe too constraining to allow their discovery. Exploring the full space restricted to 5 possible start points for each of the 21 items would require 10^{14} models. However, we have no apriori reason to expect multiple solutions outside the tolerance of this approach.

Finally, all reported estimates will be subject to some inevitable noise due to the stochastic nature of the realisations.

References

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, 34, 181-192. DOI: 10.1177/0146621609360202

Yen, W. M. (1984). Effects of LID on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement,* 8, 125-145.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, 39, 291-309.